

# The Unicode Standard

## Version 6.2 – Core Specification

To learn about the latest version of the Unicode Standard, see <http://www.unicode.org/versions/latest/>.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

Unicode and the Unicode Logo are registered trademarks of Unicode, Inc., in the United States and other countries.

The authors and publisher have taken care in the preparation of this specification, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

The *Unicode Character Database* and other files are provided as-is by Unicode, Inc. No claims are made as to fitness for any particular purpose. No warranties of any kind are expressed or implied. The recipient agrees to determine applicability of information provided.

Copyright © 1991–2012 Unicode, Inc.

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction. For information regarding permissions, inquire at <http://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <http://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 6.2.

Includes bibliographical references and index.

ISBN 978-1-936213-07-8 (<http://www.unicode.org/versions/Unicode6.2.0/>)

1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.

QA268.U545 2012

ISBN 978-1-936213-07-8

Published in Mountain View, CA

September 2012

# I General Index

The General Index covers the contents of this core specification. To find topics in the Unicode Standard Annexes, Unicode Technical Standards, and Unicode Technical Reports, use the search feature on the Unicode Web site.

For definitions of terms used, see the glossary on the Unicode Web site. To find the code points for specific characters or the code ranges for particular scripts, use the Character Index on the Unicode Web site. (See *Section B.6, Other Unicode Online Resources.*)

## A

abbreviation, Coptic . . . . . 228  
abjads . . . . . 188, 245  
abstract character sequences  
  definition . . . . . 67  
abstract characters . . . . . 21  
  definition . . . . . 67  
abugidas . . . . . 189, 190, 277, 363  
accent marks *see* diacritics  
accented characters  
  encoding . . . . . 9  
  Latin . . . . . 212  
  normalization . . . . . 152  
accounting numbers, ideographic . . . . . 134  
acrophonic numerals . . . . . 151, 226  
Aegean numbers . . . . . 473  
Afrikaans . . . . . 216  
Ainu . . . . . 429  
Aiton . . . . . 373  
Alchemical Symbols . . . . . 527  
  reference materials . . . . . 622  
Algonquian . . . . . 457  
Ali Gali . . . . . 441  
aliases  
  character name . . . . . 66, 136, 574  
  property . . . . . 123  
  property value . . . . . 123  
allocation areas . . . . . 34  
allocation of encoded characters . . . . . 33–40, 601  
Alphabetic (informative property) . . . . . 140  
alphabets . . . . . 188  
  European . . . . . 211–243  
  mathematical . . . . . 498–502  
Alpine . . . . . 468  
alternate format characters (deprecated) . . 141, 554–555  
Amharic . . . . . 438  
Ancient Symbols . . . . . 530  
angle brackets (U+2329 and U+232A)  
  deprecated for technical publication . . . . . 518  
Annexes, Unicode Standard (UAX) . . . . . xxvii, 586  
  as components of Unicode Standard . . . . . 59  
  conformance . . . . . 64  
  list of . . . . . 64  
annotation characters . . . . . 563–564  
  use in plain text discouraged . . . . . 564

ANSI/ISO C  
  wchar\_t and Unicode . . . . . 148  
apostrophe (U+0027) . . . . . 200  
Arabic . . . . . 250–266  
  digits . . . . . 503  
Arabic-Indic digits . . . . . 253–254  
  signs used with . . . . . 255  
ArabicShaping.txt . . . . . 256, 260, 271  
Aramaic . . . . . 277, 324, 355, 441, 478  
archaic scripts . . . . . 465–474  
areas of the Unicode Standard . . . . . 34  
ARIB . . . . . 524  
Armenian . . . . . 232–233  
arrows . . . . . 515–516  
ASCII  
  characters with multiple semantics . . . . . 193  
  transparency of UTF-8 . . . . . 27  
  Unicode modeled on . . . . . 1  
  zero extension . . . . . 148, 597  
Assamese . . . . . 296  
assigned code points . . . . . 8, 23  
Athapascan . . . . . 457  
atomic character boundaries . . . . . 159  
Avestan . . . . . 482  
  reference materials . . . . . 622

## B

Balinese . . . . . 394–399  
  reference materials . . . . . 623  
Bamum . . . . . 455–456  
  reference materials . . . . . 623  
Bangla . . . . . 295–299  
base characters . . . . . 238  
  definition . . . . . 79  
  multiple . . . . . 45  
  ordered before combining marks . . . . . 161, 238  
Basic Multilingual Plane (BMP) . . . . . 1, 33  
  allocation areas . . . . . 37  
  representation in UTF-16 . . . . . 27  
Basque . . . . . 216  
Batak . . . . . 402–403  
  reference materials . . . . . 623  
benefits of Unicode . . . . . 1  
Bengali . . . . . 295–299  
Bidi Class (normative property) . . . . . 130  
Bidi Mirrored (normative property) . . . . . 135  
Bidi Mirroring Glyph (informative property) . . . . 135

- BidiMirroring.txt .....135
  - Bidirectional Algorithm, Unicode .....40, 63
  - bidirectional ordering .....15
    - controls .....141, 553
  - bidirectional text .....40, 63
    - Middle Eastern scripts .....245
    - nonspacing marks in .....163
    - punctuation in .....192
  - big-endian .....30
    - definition .....63
  - Bihari .....293
  - binary comparison and sort order
    - caution for UTF-16 .....27
    - UTF differences .....168, 170
    - UTF-8 .....29
  - blocks of the Unicode Standard .....34, 187
  - Blocks.txt .....34
  - BMP *see* Basic Multilingual Plane
  - BNF (Backus-Naur Form) .....583
  - BOCU-1 *see* UTXN #6, BOCU-1
    - MIME-Compatible Unicode Compression
  - Bodhi .....325
  - Bodo .....292
  - BOM (U+FEFF) .....30, 50, 98–101, 561–562
  - Bopomofo .....426–427
  - boundaries, text .....8, 46, 140, 158–159, 167
    - see also* UAX #14, Unicode Line Breaking Algorithm
    - rhythm
    - see also* UAX #29, Unicode Text Segmentation
  - boustrophedon .....41, 470
  - Brahmi .....277, 324, 355, 359–361, 364
    - reference materials .....623
  - Braille .....534–536
  - Breton .....216
  - Buginese .....393–394
  - Buhid .....392
  - Bulgarian .....230
  - bullets .....202
    - numeric .....505
  - Burmese *see* Myanmar
  - Byelorussian .....230
  - byte order mark (BOM) (U+FEFF) ..30, 50, 98–101, 561–562
  - byte ordering
    - changing .....61
    - conformance .....62
  - byte serialization .....30, 50
  - Byzantine Musical Symbols .....540
- C**
- C language
    - wchar\_t and Unicode .....148
  - C0 and C1 control codes .....23, 139, 544
  - Cambodian *see* Khmer
  - camelcase .....174
  - Canadian Aboriginal Syllabics .....457–458
    - reference materials .....624
  - candrabindu .....294, 349
  - canonical composite characters
    - see* canonical decomposable characters
  - canonical composition algorithm .....105
  - canonical decomposable characters
    - definition .....88
  - canonical decomposition .....48
    - definition .....88
    - mappings .....87
  - canonical equivalence
    - definition .....88
    - nonspacing marks .....164
  - canonical equivalent character sequences
    - conformance .....60, 61
  - canonical mappings
    - see* canonical decomposition mappings
  - canonical ordering algorithm .....104
  - canonical precomposed characters
    - see* canonical decomposable characters
  - Cantonese .....413
  - capital letters .....124, 172, 211
  - Carian .....474
    - reference materials .....624
  - carriage return (U+000D) (CR) .....154, 545
  - carriage return and line feed (CRLF) .....154
  - case .....217
    - and text processes .....9
    - beyond ASCII .....172
    - camelcase .....174
    - case folding .....175
    - case operations (conformance) .....64, 115–120
    - case operations and normalization .....177
    - case operations, reversibility .....174
    - cased (definition) .....115
    - case-insensitive comparison ...119, 168, 169, 175
    - casing context (definition) .....116
    - conversion .....117
    - detection .....118
    - European alphabets .....211
    - exceptional Latin pairs .....214, 217
    - Georgian .....234
    - lowercase .....124, 172, 211
    - mapping tables .....146
    - mappings .....115, 126, 172–174
    - mappings noted in code charts .....576
    - titlecase .....124, 172
    - Turkish I .....173, 174, 214
    - uppercase .....124, 172, 211
    - see also* default case
  - Case (normative property) .....124, 172
  - CaseFolding.txt .....126, 175, 176
  - caseless letters .....217
  - Catalan .....215
  - cedilla .....213
  - CEF *see* character encoding forms
  - CES *see* character encoding schemes
  - CESU-8
    - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
  - Chakma .....350–351
    - reference materials .....624
  - Cham .....390–391
    - reference materials .....624
  - character encoding forms (CEF) .....24–29, 597
    - see also* Unicode encoding forms
  - character encoding model .....24, 31
    - see also* UTR #17, Unicode Character Encoding Model
  - character encoding schemes (CES) .....30–32
    - see also* Unicode encoding schemes

- character encoding standards
  - coverage by Unicode ..... 2
- Character Index ..... 589
- character literals, Unicode
  - code point notation U+ ..... 583
- character mapping
  - interchange format *see* UTS #22, Character Mapping Markup Language (CharMapML)
- character names ..... 66, 135–139, 599
  - aliases ..... 66, 136, 574
  - conventions ..... 581
  - for CJK ideographs ..... 577
  - for control codes ..... 138, 139
  - in code charts ..... 571–575
  - matching ..... 136
- character properties
  - see* properties
  - see also individual properties, e.g.* Combining Class
- character semantics ..... 1, 60, 65–66, 599
  - as Unicode design principle ..... 14
  - ASCII ..... 193
  - definition ..... 65
- character sequences
  - abstract *see* abstract character sequences
  - canonical equivalent *see* canonical equivalent character sequences
  - compatibility equivalent *see* compatibility equivalent character sequences
  - conformance ..... 60
  - named ..... 136
- character sequences, combining ..... 79
- character shaping selectors (deprecated) ..... 555
- character tabulation (U+0009) ..... 545
- characters
  - abstract *see* abstract characters
  - arrangement in Unicode ..... 35
  - assigned ..... 8, 23
  - blocks ..... 34, 187
  - boundaries ..... 158
  - canonical decomposable *see* canonical decomposable characters
  - classes ..... 583
  - code charts ..... 571–580, 589
  - coded *see* encoded characters
  - combining *see* combining characters
  - compatibility decomposable *see* compatibility decomposable characters
  - composite *see* decomposable characters
  - concept of ..... 11, 46
  - conformance definitions ..... 67–69
  - confusable ..... 179
  - conversion ..... 145–147
  - decomposable *see* decomposable characters
  - deprecated *see* deprecated characters
  - encoded *see* encoded characters
  - encoding forms *see* encoding forms
  - encoding schemes *see* encoding schemes
  - end-user perceived ..... 46
  - format control ..... 23, 51, 193, 543–569
  - glyphs, relationship to ..... 11
  - graphic ..... 23
  - identity (definition) ..... 65
  - ignored in processing ..... 180–185
  - interpretation ..... 59
  - layout control ..... 51, 545–554
  - modification ..... 61
  - names list ..... 571–575
  - names *see* character names
  - not encoded in Unicode ..... 2
  - number encoded in this and earlier versions ..... 601
  - number encoded in Version 6.2 ..... 2
  - precomposed *see* decomposable characters
  - properties *see* properties
  - semantics *see* character semantics
  - special ..... 50, 543–569
  - supplementary *see* supplementary characters
  - transcoding ..... 145–147
  - unsupported ..... 148–149
- characters, not glyphs
  - in spoofing ..... 180
  - Unicode principle ..... 11
- CharMapML
  - see* UTS #22, Character Mapping Markup Language (CharMapML)
- charsets
  - IANA registered names ..... 31
- charts, character code *see* code charts
- Cherokee ..... 456
  - reference materials ..... 624
- Chinese ..... 413–414
  - Cantonese ..... 413
  - Hakka ..... 427
  - Mandarin ..... 413
  - Minnan (Hokkien/Fujian, incl. Taiwanese) ..... 427
  - simplified and traditional ..... 413
- Chu hán ..... 412
- Chu Nôm ..... 607
- citations for
  - properties ..... 58
  - Unicode algorithms ..... 58
  - Unicode Standard ..... 57
- CJK ideographs ..... 190, 406–421
  - accounting numbers ..... 134
  - CJK Compatibility Ideographs ..... 420–421
  - CJK Compatibility Supplement ..... 421
  - CJK Strokes ..... 423, 609
  - CJK Unified Ideographs ..... 406–420
  - CJK Unified Ideographs Extension A ..... 410
  - CJK Unified Ideographs Extension B ..... 420
  - CJK Unified Ideographs Extension C ..... 420
  - CJK Unified Ideographs Extension D ..... 420
  - code charts ..... 577
  - compatibility ideographs in Plane 2 ..... 40
  - component structure ..... 416
  - encoding blocks ..... 409
  - ideographic description sequences ..... 423–425
  - ideographic variation mark (U+303E) ..... 425
  - KangXi radicals ..... 419, 421–422
  - names ..... 577
  - numbers ..... 503
  - numeric values ..... 133, 151
  - order of encoding ..... 418
  - radicals ..... 421–422
  - source standards ..... 407–409
  - unknown or unavailable ..... 208
  - Vietnamese ..... 405
- CJK Miscellaneous Area ..... 38
- CJK punctuation and symbols ..... 207
  - compatibility forms ..... 209
  - overscores and underscores ..... 209

- quotation marks .....198
- sesame dots .....208
- vertical forms .....209
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) .....606
- CJKV Ideographs Area .....38
- CLDR (Unicode Common Locale Data Repository) . . .590
- cluster boundaries .....158
- code charts ..... 571–580, 589
  - representative glyphs .....572
- code point sequences
  - notation .....582
- code points .....5, 22
  - assigned .....8, 23
  - assignment .....35, 601
  - categories .....22
  - default ignorable .....149, 184
  - definition .....67
  - designated .....23
  - notation .....581
  - number in Unicode Standard .....1
  - private-use *see* private-use code points
  - reserved *see* reserved code points
  - semantics .....24
  - surrogate *see* surrogates
  - unassigned *see* unassigned code points
  - undesignated .....23
- code positions *see* code points
- code set independence .....14
- code unit sequences
  - definition .....90
  - ill-formed (definition) .....91
  - notation .....582
  - well-formed (definition) .....91
- code units
  - definition .....90
  - isolated .....89
- code values *see* code units
- coded character representations
  - see* coded character sequences
- coded character sequences
  - definition .....68
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng .....375, 376
- Collation Algorithm, Unicode (UCA) .....10
- collation *see* sorting
- collation tables .....146
- combining character sequences .....42, 79
  - defective .....163
  - definition .....81
  - Latin .....212
  - line breaking .....160
  - matching .....160
  - order of base character and marks .....161, 238
  - rendering .....160
  - selection .....158
  - truncation ..... 161–162
- combining characters ..... 41–46, 83–86, 159–166
  - blocking reordering .....551
  - canonical ordering ..... 47, 104, 126
  - class zero .....127
  - combining marks ..... 238–239
  - definition .....80
  - dependence .....238
  - display order .....43
  - keyboard input .....160
  - ligatures .....45
  - multiple .....43
  - multiple base characters .....45
  - normalization of .....152
  - ordering conventions .....42
  - rendering of marks .....162–166
  - reordrant .....127
  - script-specific .....42
  - split .....128
  - strikethrough .....130
  - subjoined .....129
  - typographical interaction .....43, 126
  - vertical stacking .....44
  - see also* diacritics
- Combining Class (normative property) .....126
- combining classes .....102, 126, 165
  - class zero characters .....126
  - definition .....102
- combining grapheme joiner (U+034F) .....551
- combining half marks .....141, 243
- combining marks *see* combining characters
- comma below .....213
- Compatibility and Specials Area .....20, 38
- compatibility characters .....18
- compatibility composite characters .....21
  - see* compatibility decomposable characters
- compatibility decomposable characters .....20
  - definition .....87
- compatibility decomposition .....48
  - definition .....87
- compatibility decomposition mappings .....87
- Compatibility Encoding Scheme for UTF-16
  - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
- compatibility equivalence
  - definition .....88
- compatibility equivalent character sequences
  - conformance .....61
- compatibility mappings
  - see* compatibility decomposition mappings
- compatibility precomposed characters
  - see* compatibility decomposable characters
- compatibility variants .....20
  - mapping .....177
- composite characters
  - see* decomposable characters
  - compatibility *see* compatibility decomposable characters
- Composition Exclusion (normative property) . . .74
- compression .....153
  - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
- conferences .....589
- conformance .....55–120
  - clause and definition updates .....603
  - definitions .....65–69
  - examples .....51
  - ISO/IEC 10646 implementations .....600
  - requirements .....59–63
- confusables .....179
- conjunct consonants
  - Indic .....158, 281

Myanmar	369	definition	87
selection of clusters	158	in normalization	152
contextual shaping		mapping, definition	87
apostrophe	200	mappings noted in code charts	576
Arabic	251	default case	
not used for Hebrew final forms	247	algorithms	64, 115–120
quotation marks	197	conversion	117
Syriac	269	detection	118
contour tones	237	folding	117
control codes	23, 51, 544	default caseless matching	119
graphics for	517	default grapheme clusters	159
names	139	<i>see also</i> UAX #29, Unicode Text Segmentation	
properties	545	Default Ignorable Code Point (property)	184
semantics	24, 545	default ignorable code points	149, 184
specified in Unicode	545	default property values	72
control sequences	544	definition	72
conversion of characters	96, 145–147, 185	defective combining character sequences	163
convertibility		definition	81
as Unicode design principle	19	dependent vowel signs	
Coptic	225, 227–229	Indic	280
reference materials	624	Khmer	378
corporate use subarea	558	Philippine scripts	392
corrigenda	57	deprecated characters	55, 574
CR (U+000D carriage return)	154, 545	alternate format	141, 554–555
CRLF (carriage return and line feed)	154	definition	68
Croatian	216	Derived Age (property)	149
digraphs	216	derived properties	
culturally expected sorting	10, 168	definition	78
Cuneiform		DerivedCoreProperties.txt	116, 124, 184
Old Persian	484	DerivedNormalizationProps.txt	177
Sumero-Akkadian	485–487	Deseret	459–460
Ugaritic	483	reference materials	625
Cuneiform and Hieroglyphic Area	39	design goals of Unicode	3
currency symbols	494–496	design principles of Unicode	10–19
encoded in script blocks	495	designated code points	23
cursive joining	548–551	Devanagari	278–295
Arabic	256–262	Dhivehi	274
control characters for	141, 252, 443, 547	diacritics	42, 238
Mandaic	480	alternative glyphs	212, 239
Mongolian	442–444	Czech	212
N’Ko	453	display in isolation	45, 195, 239
Syriac	269–272	double	85, 141, 240
transparency	550	Greek	222–223, 226
cursive scripts	245	Latin	212–214
Cypriot	474	Latvian	213
reference materials	629	mathematical	501
<i>see also</i> Linear B		on i and j	214
Cyrillic	229–231	rendering	162–166
Czech	216	Slovak	212
		spacing clones of	237, 239
<b>D</b>		symbol	42, 242
danda, in Devanagari block	291	<i>see also</i> combining characters	
Danish	215	dictionary symbols	524
dashes	195	digit form names	254
Database, Unicode Character		digits	151
<i>see</i> Unicode Character Database (UCD)		Arabic	503
dead consonants, Indic	281	Arabic-Indic	253–254
dead keys	160	compatibility	504
decomposable characters	48	decimal	133
definition	87	glyph variants	505
normalization of	152	hexadecimal	504
decomposition	48, 87–89	Myanmar	503
canonical <i>see</i> canonical decomposition		national shapes	555
compatibility <i>see</i> compatibility decomposition		Shan	503
		superscript and subscript	504

Tai Tham .....503  
 digraphs ..... 216, 219, 220  
 dingbats ..... 526–527  
 directionality .....15, 40  
 East Asian scripts .....406  
 Middle Eastern scripts .....245  
 Mongolian .....442  
 musical symbols .....537  
 normative property .....130  
 Ogham .....467  
 Old Italic .....468  
 Philippine scripts .....393  
 Runic .....470  
 discussion list for Unicode .....590  
 Dogri .....292  
 Domino Tiles .....528  
 dotless i ..... 173, 174, 214  
 dotted circle  
 in code charts .....80, 239  
 in fallback rendering .....163  
 to indicate diacritic .....41  
 to indicate vowel sign placement .....43  
 double diacritics ..... 85, 141, 240  
 Dutch .....215, 216  
 dynamic composition  
 as Unicode design principle .....18  
 Dzongkha .....325

## E

East Asian scripts ..... 405–435  
 writing direction .....41  
*see also* CJK ideographs  
 Eastern Arabic-Indic digits .....253  
 EBCDIC  
 newline function .....154  
*see* UTR #16, UTF-EBCDIC  
 editing, text boundaries for ..... 158–159  
 efficiency  
 as Unicode design principle .....11  
 Egyptian hieroglyphs ..... 487–491  
 reference materials .....625  
 e-mail discussion list for Unicode .....590  
 emoji .....522, 523  
 animal symbols .....525  
 cultural symbols .....525  
 zodiacal symbols .....525  
 Emoticons .....525  
 Enclosed Alphanumerics .....533  
 enclosing marks .....243  
 definition .....80  
 encoded characters .....5, 22  
 allocation ..... 33–40, 601  
 definition .....67  
 encoding form conversion  
 definition .....95  
 encoding forms .....24–29  
 ISO/IEC 10646 definitions .....597  
 encoding forms, Unicode  
*see* Unicode encoding forms  
 encoding model for Unicode characters .....24, 31  
*see also* UTR #17, Unicode Character Encoding Model  
 encoding schemes .....30–32

encoding schemes, Unicode  
*see* Unicode encoding schemes  
 endian ordering  
*see* byte order mark (BOM) (U+FEFF)  
 end-user subarea .....559  
 English .....215  
 equivalent sequences .....152  
 as Unicode design principle .....18  
 case-insensitivity .....169, 175  
 combining characters in matching .....160  
 conformance .....61  
 Hangul syllables .....432  
 in sorting and searching .....167  
 language-specific .....88  
 security implications .....179  
*see also* canonical equivalence  
*see also* compatibility equivalence  
*see also* encoding forms, encoding schemes  
 errata ..... xxviii, 57, 590  
 escape sequences .....544  
 not used in Unicode .....1, 3  
 Esperanto .....216  
 Estonian .....216  
 Ethiopic .....438–440  
 reference materials .....625  
 Etruscan .....467  
 euro sign (U+20AC) .....496  
 European alphabetic scripts .....211–243  
 eyelash-RA .....286

## F

fallback rendering .....184  
 of nonspacing marks .....162  
 FAQ (Frequently Asked Questions) .....590  
 Faroese .....215  
 Farsi .....250, 252  
 featural syllabaries .....189  
 FF (U+000C form feed) .....154, 545  
 file separator (U+001C) .....545  
 Finnish .....215  
 Finno-Ugric Transcription (FUT)  
*see* Uralic Phonetic Alphabet (UPA)  
 fixed-width Unicode encoding form (UTF-32) 26, 93  
 flat tables .....146  
 Flemish .....215  
 fonts  
 and Unicode characters .....13  
 for mathematical alphabets .....500–502  
 style variation for symbols .....493  
 form feed (U+000C) (FF) .....154, 545  
 format control characters .....23, 51, 193, 543–569  
 deprecated .....554–555  
 prefixed .....141  
 stateful .....553  
 fraction characters .....511  
 fraction slash (U+2044) .....200, 509  
 French .....216  
 Frisian .....216  
 FTP site, Unicode Consortium .....589  
 fullwidth forms in East Asian encodings .....429–430  
 futhark .....470

## G

Garshuni .....266

- Ge'ez ..... 438
- General Category (normative property) ..... 130  
list of values ..... 130
- general punctuation ..... 191–209
- General Scripts Area ..... 38
- geometrical symbols ..... 520–522
- Georgian ..... 233–235
- German ..... 215
- geta mark (U+3013) ..... 208
- Glagolitic ..... 231–232  
reference materials ..... 625
- Glossary ..... 590
- glyph selection tables ..... 146
- glyphs ..... 5, 12  
characters, relationship to ..... 11  
diacritics alternative ..... 212, 239  
Greek alternative ..... 223–224  
Latin alternative ..... 212  
mathematical alternative ..... 512  
missing ..... 184  
representative in code charts ..... 572  
standardized variants ..... 556  
symbols alternative ..... 493
- golden numbers ..... 471
- Gothic ..... 471–472  
reference materials ..... 626
- grapheme base ..... 238  
definition ..... 81
- grapheme clusters ..... 8, 46  
*see also* UAX #29, Unicode Text Segmentation  
default ..... 159  
definition ..... 82
- grapheme extender  
definition ..... 81
- grapheme joiner, combining (U+034F) ..... 551
- graphic characters ..... 23
- Greek ..... 222–226  
acrophonic numerals ..... 151, 226  
alternative glyphs ..... 223–224  
ancient musical notation ..... 540–542  
editorial marks ..... 205, 626  
letters as symbols ..... 223–225, 513  
*see also* Cypriot, Linear B
- Greenlandic ..... 216
- group separator (U+001D) ..... 545
- guillemets ..... 198
- Gujarati ..... 303–304
- Gurmukhi ..... 300–303
- ## H
- Hakka ..... 427
- halant ..... 277  
*see also* virama
- half marks, combining ..... 141, 243
- half-consonants, Indic ..... 282
- halfwidth forms in East Asian encodings ... 429–430
- Han ideographs *see* CJK ideographs
- Han unification ..... 414–420  
and language tags ..... 157  
history ..... 605–607  
language usage ..... 412  
source separation rule ..... 410, 415  
source standards ..... 407–409
- Hangul Area ..... 38
- Hangul syllables ..... 405, 430–433  
and combining marks ..... 86  
as grapheme clusters ..... 46  
canonical decomposition ..... 109  
collation ..... 433  
composition ..... 111  
conjoining jamo ..... 107–115  
equivalent sequences ..... 432  
Hangul Compatibility Jamo ..... 431  
Hangul Jamo ..... 430–433  
Hangul Syllables block ..... 432–433  
Johab set ..... 432  
name generation ..... 112  
normalization ..... 431  
standard ..... 108
- Hangzhou numerals ..... 508
- Hanja *see* CJK ideographs
- Hanunóo ..... 392
- Hanzi *see* CJK ideographs
- harakat, Arabic pronunciation marks ..... 250
- hasant ..... 296
- hash tables ..... 146
- Hebrew ..... 246–250
- hentaigana ..... 429
- hieroglyphs  
Egyptian ..... 487–491  
Meroitic ..... 491–492
- high surrogate  
definition ..... 89  
high-surrogate code points ..... 59, 559  
high-surrogate code units ..... 89
- higher-level protocols  
definition ..... 69
- Hindi ..... 278
- Hiragana ..... 428
- historic scripts ..... 465–474
- horizontal tab (U+0009) ..... 545
- HTML newline function ..... 155
- Hungarian ..... 216
- hyphenation ..... 547  
as a text process ..... 8
- hyphens ..... 195, 547
- ## I
- I Ching symbols ..... 529
- IANA charset names ..... 31
- Icelandic ..... 215
- identifiers ..... 167  
*see also* UAX #31, Unicode Identifier and Pattern  
Syntax
- Ideographic (informative property) ..... 140
- ideographic description sequences ..... 424
- Ideographic Rapporteur Group (IRG) ..... 407, 606
- Ideographic Variation Database *see* UTS #37, Unicode  
Ideographic Variation Database
- ideographs *see also* CJK ideographs
- IDNA *see* UTS #46, Unicode IDNA Compatibility  
Processing
- IICore ..... 411, 607
- ill-formed  
definition ..... 91
- Imperial Aramaic ..... 478–479  
reference materials ..... 626
- implementation guidelines ..... 145–185



- in a Unicode encoding form
    - definition .....92
  - in-band mechanisms .....568
  - Indian rupee sign (U+20B9) .....496
  - Indic scripts ..... 277–322, 323–325
    - principles, in terms of Devanagari ..... 279–285
    - relation to ISCII standard .....278
  - Indonesian .....215
  - industry character sets
    - covered in Unicode .....2
  - information separators (U+001C..U+001F) ....545
  - informative properties
    - definition .....75
  - Inscriptional Pahlavi .....481
  - Inscriptional Parthian .....481
  - inside-out rule .....162
  - interchange restrictions .....23
  - International Phonetic Alphabet (IPA) 188, 218–219
    - reference materials .....627
    - Spacing Modifier Letters .....236
    - see also* phonetic alphabets
  - internationalization .....14
  - Internationalization & Unicode Conference ....589
  - Internet protocols
    - UTF-8 as preferred encoding .....28
  - Inuktitut .....457
  - invisible operators .....516
  - iota subscript .....223
  - IPA *see* International Phonetic Alphabet
  - IRG (Ideographic Rapporteur Group) .....407, 606
  - Irish .....215, 466
  - ISCII standard and Unicode .....278
  - ISO/IEC 10646 ..... 593–600
    - conformance of Unicode implementations ...599
    - encoding forms .....597
    - synchrony with Unicode Standard .....598
    - timeline compared to Unicode versions ....594
  - Italian .....215
  - ITC Zapf Dingbats .....526
  - IUC *see* Internationalization & Unicode Conference
- J**
- jamos *see* Hangul syllables
  - Japanese .....405
  - Javanese ..... 399–401
    - reference materials .....627
  - Jawi .....264
  - jihvamuliya .....295, 349
  - Johab .....432
  - joiners .....252
    - combining grapheme joiner (U+034F) .....551
    - word joiner (U+2060) .....546
    - zero width joiner (U+200D) .....252, 549
  - justification .....164
- K**
- Kaithi ..... 345–347
    - reference materials .....627
  - Kana (Hiragana and Katakana) ..... 428–429
  - Kanbun .....421
  - KangXi radicals ..... 419, 421–422
  - Kanji *see* CJK ideographs
  - Kannada ..... 315–317
  - Kashmiri .....293
  - Katakana .....428–429
  - Kawi .....394, 396
  - Kayah Li ..... 389–390
    - reference materials .....627
  - KC (normalization form)
    - see* Normalization Form KC
  - KD (normalization form)
    - see* Normalization Form KD
  - keytop labels .....517
  - Khamti Shan .....372
  - Kharoshthi .....355–356
    - reference materials .....628
  - Khmer .....374–383
    - characters not recommended .....380
    - syllable components, order of .....381
  - killer .....189
    - Batak .....402
    - Brahmi .....359
    - Meetei Mayek .....352
    - Myanmar (asat) .....370
    - see also* virama
  - Konkani .....292
  - Korean Hangul *see* Hangul
  - Kurdish .....250, 264
- L**
- Ladino .....246
  - language tags .....157, 565–568
    - and Han unification .....157
    - use strongly discouraged .....568
  - Lanna .....384
  - Lao .....366–368
  - last-resort glyphs .....184
  - Latin .....212–222
    - alternative glyphs .....212
    - Basic Latin .....215
    - encoding blocks .....34
    - IPA Extensions .....218–219
    - Latin Extended Additional .....220–222
    - Latin Extended-A .....216
    - Latin Extended-B .....216–218
    - Latin Extended-C .....220
    - Latin Extended-D .....221
    - Latin Ligatures .....220
    - Latin-1 Supplement .....215
    - Phonetic Extensions .....219–221
  - Latvian .....216, 221
    - cedilla .....213
  - layout control characters .....51, 545–554
  - leading surrogates
    - see* high-surrogate code units
  - legibility criterion for plain text .....15
  - Lepcha .....335–336
    - reference materials .....628
  - letter spacing .....547
  - letterlike symbols .....496–502
  - LF (U+000A line feed) .....154, 545
  - ligatures .....548–551
    - Arabic .....258–259
    - combining characters on .....45
    - control characters for .....141
    - for nonspacing marks .....165
    - Latin .....220
    - selection .....159

Syriac	272	mathematical expression format characters	141
Limbu	342–344	<i>see also</i> UTR #25, Unicode Support for Mathematics	
reference materials	628	mathematical symbols	511–516
line breaking	153–156, 546–548	alphabets	498–502
control characters	143	alphanumeric	498–502
in South Asian scripts	366, 371, 383	fonts	500–502
recommendations	155	format characters	516
<i>see also</i> UAX #14, Unicode Line Breaking Algorithm		fragments for typesetting	518
line feed (U+000A) (LF)	154, 545	invisible operators	516
line separator (U+2028) (LS)	154, 547	operators	512–513
line tabulation (U+000B) (VT)	545	reference materials	630
Linear B	473	standardized variants	516
reference materials	629	MathML	513
<i>see also</i> Cypriot		matras	127, 280
linear boundaries	159	Meetei Mayek	351–353
lira sign, Turkish (U+20BA)	496	reference materials	630
Lisu	461–463	Meroitic	
reference materials	629	cursive	491–492
Lithuanian	216	hieroglyphs	491–492
little-endian	30	reference materials	630
definition	63	Miao	463–464
Locale Data Markup Language		reference materials	631
<i>see</i> UTS #35, Unicode Locale Data Markup Language (LDML)		Middle Eastern scripts	245–275
logical order		Min	413
as Unicode design principle	15	Minnan (Hokkien/Fujian, incl. Taiwanese)	427
exceptions to	128	minor version	56
logograph	190	minus sign	513
logosyllabaries	190	commercial (U+2052)	203
low surrogate		mirrored property	
definition	89	<i>see</i> Bidi Mirrored (normative property)	
low-surrogate code points	59, 559	mirroring of paired punctuation	197
low-surrogate code units	89	Miscellaneous Symbols	523
lowercase	124, 172, 211	missing glyphs	184
LS (U+2028 line separator)	154, 547	modifier letters	235–238
Lycian	474	Modifier Letters, Spacing	220
reference materials	630	Mongolian	337, 440–447
Lydian	474	writing direction	442
reference materials	630	multibyte encodings	
		compared to UTF-8	28
<b>M</b>		multistage tables	146
MacOS newline function	154	musical symbols	536–542
Mahjong Tiles	527	ancient Greek	540–542
mail discussion list for Unicode	590	Balinese	398
Maithili	292	Byzantine	540
major version	56	directionality	537
Malay	215	Gregorian	537
Malayalam	317–322	reference materials	631
Maltese	216	Western	536–539
Manchu	441	Myanmar	368–373
Mandaic	479–481	digits	503
reference materials	630	Myanmar Extended-A	371
Mandarin	413	reference materials	632
Manden	450	<b>N</b>	
map symbols	524	N’Ko	450–454
mapping tables <i>see</i> tables of character data		reference materials	632
Marathi	278, 286, 290	named character sequences	136
markup languages		names, character <i>see</i> character names	
and Unicode conformance	568	namespace	66
line breaking	153	NEL (U+0085 next line)	154, 545
<i>see also</i> UTR #20, Unicode in XML and Other Markup Languages		Nepali	278
Mathematical (informative property)	511	neutral directional characters	130
		New Tai Lue	384–385

- newline function (NLF) .....154, 545  
 newline guidelines ..... 153–156  
 next line (U+0085) (NEL) .....154, 545  
 NFC (Normalization Form C) .....47  
 NFD (Normalization Form D) .....47  
 NFKC (Normalization Form KC) .....47  
 NFKD (Normalization Form KD) .....47  
 NLF (newline function) .....154, 545  
 no-break space (U+00A0) .....546  
     base for diacritic in isolation ..... 45, 195, 239  
 no-break space, narrow (U+202F) .....445  
 noncharacter code points *see* noncharacters  
 noncharacters .....23, 50, 560  
     conformance .....59  
     definition .....68  
     handling .....61  
     in code charts .....574  
     interchange restrictions .....24  
     semantics .....24  
     U+10FFFF (not a character code) .....560  
     U+FDD0..U+FDEF .....23, 560  
     U+FFFE (not a character code) .....50, 560  
     U+FFFF (not a character code) .....23, 560  
 nondecomposable characters .....48  
 non-joiner, zero width (U+200C) .....252, 549  
 nonlinear boundaries .....159  
 non-overlap principle in Unicode encoding forms .24  
 nonspacing marks .....238  
     definition .....80  
     display in isolation ..... 45, 195, 239  
     positioning .....165  
     rendering ..... 162–166  
     *see also* combining characters  
     *see also* diacritics  
 normalization .....47, 152  
     and case operations .....177  
     canonical ordering algorithm ..... 47, 104, 126  
     conformance .....63  
     of private-use characters .....558  
     *see also* UAX #15, Unicode Normalization Forms  
     stability .....102  
 Normalization Form C (NFC) .....47  
 Normalization Form D (NFD) .....47  
 Normalization Form KC (NFKC) .....47  
 Normalization Form KD (NFKD) .....47  
 normalization forms ..... 101–107  
     definition .....107  
     specification .....103  
 normative behaviors  
     definition .....65  
 normative properties  
     definition .....74  
     list .....74  
     may change .....74  
 Norwegian .....215  
 notational conventions ..... 581–584  
 notational systems .....191  
 nukta .....265, 287  
 null (U+0000)  
     as Unicode string terminator .....545  
 number forms  
     CJK ideographs .....151  
 numbers  
     handling .....151  
     ideographic accounting .....134  
 numerals .....502–510  
     acrophonic .....226  
     Chinese counting rods .....510  
     Coptic .....229  
     Cuneiform .....487  
     Ethiopic .....439  
     Greek acrophonic .....151  
     Hangzhou .....508  
     old-style .....201  
     Roman .....151, 511  
     Rumi .....507  
     Suzhou-style .....508  
 numeric separators .....203  
 numeric shape selectors (deprecated) .....555  
 Numeric Type (normative property) .....133  
 Numeric Value (normative property) .....133  
 numero sign (U+2116) .....496
- O**
- object replacement character (U+FFFC) .....564  
 octet .....583  
 Ogham .....466–467  
     reference materials .....632  
 Ol Chiki .....353–354  
     reference materials .....632  
 Old Italic .....467–469  
     reference materials .....632  
 Old Persian .....484–485  
     reference materials .....633  
 Old South Arabian .....475–477  
     reference materials .....633  
 Old Turkic .....472  
     reference materials .....633  
 old-style numerals .....201  
 Oriya .....304–306  
 Oromo .....438  
 Osmanya .....447  
     reference materials .....633  
 out-of-band mechanisms .....568  
 overlapping encodings .....24  
 overscores .....201
- P**
- Pahlavi, Inscriptional .....481  
     reference materials .....626  
 Panjabi .....300  
 paragraph or section marks .....203  
 paragraph separator (U+2029) (PS) .....154, 547  
 Parthian, Inscriptional .....481  
     reference materials .....626  
 Pashto .....250  
 Persian .....250, 252  
 Phags-pa .....336–341  
     reference materials .....634  
 Phaistos Disc symbols .....530  
 Phake .....373  
 Philippine scripts .....392–393  
     reference materials .....634  
 Phoenician .....477  
     reference materials .....634  
 phonemes .....190  
 phonetic alphabets .....188  
     IPA Extensions .....218–219  
     Phonetic Extensions .....219–221

Spacing Modifier Letters . . . . . 236–238  
 Uralic Phonetic Alphabet (UPA) . . . . . 203, 219  
*see also* International Phonetic Alphabet (IPA)  
 Pinyin . . . . . 215  
 pivot code, Unicode as . . . . . 146  
 plain text  
   as Unicode design principle . . . . . 14  
   legibility criterion . . . . . 15  
 planes of Unicode codespace . . . . . 33  
   Plane 0 (BMP) . . . . . 33  
   Plane 1 (SMP) . . . . . 33, 39  
   Plane 14 (SSP) . . . . . 33  
   Plane 2 (SIP) . . . . . 33, 40  
   Planes 15-16 (Private Use) . . . . . 40, 559  
 Playing Cards . . . . . 528  
 points, Hebrew pronunciation marks . . . . . 246  
 policies of the Unicode Consortium . . . . . 590  
 Polish . . . . . 216  
 Portuguese . . . . . 215  
 precomposed characters  
   *see* decomposable characters  
   compatibility *see* compatibility decomposable  
   characters  
 prefixed format control characters . . . . . 141  
 Private Use Area (PUA) . . . . . 38, 558  
 Private Use planes . . . . . 34, 40, 559  
 private-use characters  
   properties . . . . . 557  
   semantics . . . . . 24  
 private-use code points . . . . . 23, 148  
   conformance . . . . . 60  
   definition . . . . . 79  
   high surrogates . . . . . 559  
 processing code, choice of Unicode encoding form 28  
 properties . . . . . 14, 70–79, 121–143  
   aliases . . . . . 123  
   aliases (definition) . . . . . 78  
   and Unicode algorithms . . . . . 74  
   data tables . . . . . 146  
   derived *see* derived properties  
   in Unicode Character Database (UCD) . . . . . 34  
   informative *see* informative properties  
   normative references to . . . . . 58, 63  
   normative *see* normative properties  
   of control codes . . . . . 545  
   provisional *see* provisional properties  
   simple *see* simple properties  
   *see also individual properties, e.g. combining*  
   classes  
 property values  
   aliases . . . . . 123  
   aliases (definition) . . . . . 78  
   default . . . . . 72  
   default (definition) . . . . . 72  
   normative references to . . . . . 63  
 PropertyAliases.txt . . . . . 78, 583  
 PropertyValueAliases.txt . . . . . 78, 583  
 PropList.txt . . . . . 126  
 Provençal . . . . . 216  
 provisional properties  
   definition . . . . . 76  
 PS (U+2029 paragraph separator) . . . . . 154, 547  
 PUA (Private Use Area) . . . . . 38, 558  
*pulli* . . . . . 306

punctuation . . . . . 191–209  
   blocks containing . . . . . 187  
   CJK . . . . . 207  
   doubled . . . . . 201  
   in bidirectional text . . . . . 192  
   paired . . . . . 197  
   small form variants . . . . . 209  
   typographic forms . . . . . 192  
   vertical forms . . . . . 209  
 Punctuation and Symbols Area . . . . . 38  
 Punjabi . . . . . 300

## Q

quotation marks . . . . . 197–199  
   East Asian . . . . . 199  
   European . . . . . 198

## R

radicals, KangXi and other CJK . . . . . 421–422  
 radical-stroke index . . . . . 419  
 record separator (U+001E) . . . . . 545  
 recycling symbols . . . . . 524  
 referencing . . . . . 63  
   properties . . . . . 58  
   Unicode algorithms . . . . . 58  
   Unicode Standard . . . . . 57  
 regional indicator symbols . . . . . 534  
 regular expressions . . . . . 156  
   and line breaking . . . . . 153  
   *see also* UTS #18, Unicode Regular Expressions  
 Rejang . . . . . 401–402  
   reference materials . . . . . 635  
 rendering of text . . . . . 5, 8, 13  
   fallback . . . . . 184  
   unsupported characters . . . . . 149  
 repertoire of abstract characters . . . . . 22  
 replacement character (U+FFFD) 32, 51, 62, 96, 185,  
   565  
 reserved code points . . . . . 23, 148  
   definition . . . . . 68  
   in code charts . . . . . 574  
   preservation in interchange . . . . . 24  
   *see also* unassigned code points  
 Rhaeto-Romanic . . . . . 216  
 rich text . . . . . 14  
 right single quotation mark (U+2019)  
   preferred for apostrophe . . . . . 200  
 right-to-left text . . . . . 40  
   East Asian scripts . . . . . 406  
   Middle Eastern scripts . . . . . 245  
 roadmap for script additions . . . . . 34  
 Roman numerals . . . . . 151, 511  
 Romanian . . . . . 216  
   comma below . . . . . 213  
 Romany . . . . . 216  
 Rumi numeral forms . . . . . 507  
 Runic . . . . . 469–471  
   reference materials . . . . . 635  
 rupee sign, Indian (U+20B9) . . . . . 496  
 Russian . . . . . 229

## S

Samaritan . . . . . 273–274

- reference materials .....635
- Sami .....216
- Sanskrit .....278
- Saurashtra ..... 347–348
  - reference materials .....635
- scalar values, Unicode
  - see Unicode scalar values
- scripts
  - in Unicode Standard .....2
  - roadmap for future additions .....34
  - types of .....191
  - see also UAX #24, Unicode Script Property
- SCSU
  - see UTS #6, A Standard Compression Scheme for Unicode
- searching ..... 167–169
  - as a text process .....8
  - case-insensitive .....169, 175
- section or paragraph marks .....203
- security issues .....179
- self-synchronization of encoding forms .....25
- semantics
  - see character semantics
- sequences
  - notation .....582
- Serbian
  - corresponding digraphs in Croatian .....216
- Shan .....383
  - digits .....503
- Sharada ..... 348–349
  - reference materials .....635
- Shavian .....461
  - reference materials .....635
- Show Hidden ..... 61, 163, 184, 557
- SHY (U+00AD soft hyphen) .....547
- Sibe .....442
- signature for Unicode data ..... 51, 561–562
- simple properties
  - definition .....78
- simplified Chinese .....413
- Sindhi .....250, 292
- Sinhala ..... 324–325
  - reference materials .....636
- SIP (Supplementary Ideographic Plane) .....33, 40
- slash, fraction (U+2044) .....200
- Slovak .....216
- Slovenian .....216
- small letters ..... 124, 172, 211
- SMP (Supplementary Multilingual Plane) .....33, 39
- soft hyphen (U+00AD) (SHY) .....547
- Somali .....447
- Sora Sompeng .....354
  - reference materials .....636
- Sorbian .....216
- sorting .....10, 167
  - and combining grapheme joiner .....552
  - as a text process .....8
  - case-insensitive .....168
  - culturally expected .....10, 168
  - language-insensitive .....168
  - see also Unicode Collation Algorithm (UCA)
- source separation rule .....410, 415
- South Asian scripts ..... 277–322, 323–344
- Southeast Asian scripts ..... 363–393
  - space (U+0020)
    - base for diacritic in isolation .....46, 195, 239
  - space characters .....194, 546–548
    - graphics for .....517
  - space, zero width (U+200B) .....194
  - spacing clones of diacritics .....237, 239
  - spacing marks .....238
    - definition .....81
  - Spacing Modifier Letters .....236–238
  - Spanish .....215
  - special characters .....50, 543–569
  - SpecialCasing.txt .....115, 126
  - Specials .....561–565
  - spell-checking
    - as a text process .....8
  - spellings, alternative
    - see equivalent sequences
  - spoofing .....179
- SSP (Supplementary Special-purpose Plane) .....33
- stability .....76, 122
  - as Unicode design principle .....18
- stacked boundaries .....159
- stacking sequences .....43
  - nondefault .....44
- Standard Compression Scheme for Unicode (SCSU)
  - see UTS #6, A Standard Compression Scheme for Unicode
- standardized variants .....444, 556
  - in the code charts .....577
  - mathematical symbols .....516
- StandardizedVariants.txt .....444, 516
- standards coverage .....2
- starters .....103
- stateful encoding
  - not used in Unicode .....3
  - paired format controls .....553
- string comparison .....10
- string literals, Unicode
  - code point notation `\u1234` .....583
- strings, Unicode .....32, 91
  - null termination .....545
- strong directional characters .....130
- styled text .....14
- sublinear searching .....169
- subsets, supported .....53
  - conformance .....60
  - ISO/IEC 10646 specification for .....599
- substitution character
  - see replacement character
- Sumero-Akkadian .....485–487
- Sundanese .....403–404
  - reference materials .....636
- superscripts .....236
  - and subscripts .....510
- supplementary characters
  - in UTF-16 strings .....32
  - tables for .....146
- Supplementary General Scripts Area .....38
- Supplementary Ideographic Plane (SIP) .....33, 40
- Supplementary Multilingual Plane (SMP) .....33, 39
- supplementary planes
  - representation in UTF-16 .....27
  - representation in UTF-8 .....28
- Supplementary Private Use Areas .....40, 559
- Supplementary Special-purpose Plane (SSP) .....33

supported subsets	53	Tagbanwa	392
conformance	60	tags, language	157, 565–568
supralineation	228	use strongly discouraged	568
surrogate code points		Tai Le	383–384
<i>see</i> surrogates		reference materials	636
surrogate pairs	27, 94	Tai Tham	385–387
definition	89	digits	503
processing	28, 149–151	reference materials	637
surrogates	23, 89, 559	Tai Viet	387–389
interchange restrictions	23	Tai Xuan Jing symbols	529
isolated surrogates, handling	32	Takri	349–350
isolated surrogates, ill-formed	94	reference materials	637
isolated surrogates, uninterpreted	89	Tamil	306–313
support levels	150	TCHAR in Win32 API	148
Surrogates Area	38, 559	Technical Notes (UTN)	589
Suzhou-style numerals	508	Technical Reports (UTR)	586
svasti signs	331	abstracts	587
Swahili	215	Technical Standards (UTS)	xxviii, 586
Swedish	215	abstracts	586
syllabaries	188	technical symbols	517–520
alphabetic property	140	Telugu	313–315
featural	189	terminal emulation	494
Syloti Nagri	344–345	text boundaries	8, 46, 140, 158–159, 167
symbols	493–542	<i>see also</i> UAX #14, Unicode Line Breaking Algo-	
animal	525	rithm	
appearance variation	493	<i>see also</i> UAX #29, Unicode Text Boundaries	
arrows	515–516	text elements	5, 8, 158
cultural	525	boundaries	167
currency	494–496	for sorting	168
dictionary	524	variable-width nature	29
dingbats	526–527	text processes	4, 8–10
emoji	522, 523, 534	text rendering	5, 8, 13
Enclosed Alphanumerics	533	text selection, boundaries for	158–159
fragments for mathematical typesetting	518	Thaana	274–275
game	525	reference materials	637
gender	524	Thai	364–366
genealogical	525	Tibetan	325–334
geometrical	520–522	Tifinagh	448
Khmer lunar calendar	383	Tigre	438
letterlike	496–502	tilde (U+007E)	203
map	524	titlecase	124, 172
mathematical	511–516	Todo	441
mathematical alphanumeric	498–502	tone letters	237–238
miscellaneous	523	tone marks	
musical	536–542	Bopomofo spacing	426, 427
numerals	502–510	Chinantec	238
recycling	524	Chinese	237
regional indicator	534	Tai Le	383
technical	517–520	Thai	364
weather	524	Vietnamese	214
zodiacal	525	traditional Chinese	413
symmetric swapping format characters (deprecated)	555	traffic signs	524
Syriac	266–272	trailing surrogates	
reference materials	636	<i>see</i> low-surrogate code units	
<b>T</b>		transcoding	145–147
tab (U+0009 character tabulation)	545	tables	146
tab, vertical (U+000B)	154, 545	Transport and Map Symbols	526
tables of character data	145–147	triangulation in transcoding	146
optimization	146	tries	146
supplementary characters	146	truncation	
tag characters	565–569	combining character sequences	161–162
Tagalog	392	surrogates and	151
		Turkish	216
		case mapping of I	173, 174, 214
		cedilla	213

- lira sign (U+20BA) .....496
- two-stage tables .....146
- U**
- U+ notation .....583
- U+10FFFF (not a character code) .....560
- U+FEFF (BOM) ..... 561–562
- U+FFFE (not a character code) .....560
- U+FFFF (not a character code) .....560
- UAX (Unicode Standard Annex) ..... xxvii, 586
  - as component of Unicode Standard .....59
  - conformance .....64
  - list of .....64
- UCA *see* Unicode Collation Algorithm
- UCD *see* Unicode Character Database
- UCS (Universal Character Set)
  - see* ISO/IEC 10646
- UCS-2 .....597
- UCS-4 .....597
- Ugaritic ..... 483–484
  - reference materials .....637
- Uighur .....337, 441
- Ukrainian .....230
- unassigned code points .....23, 59, 149
  - defined as reserved code points .....68
  - handling .....55
  - properties of .....72
  - semantics .....59
  - see also* reserved code points
- underscores .....201
- undesignated code points .....23
- Unicode 1.0 Name (informative property) .....139
- Unicode algorithms
  - and properties .....74
  - conformance .....63
  - definition .....69
  - normative references to .....58, 63
- Unicode Bidirectional Algorithm .....16, 40
  - see also* UAX #9, Unicode Bidirectional Algorithm
- Unicode Character Database (UCD) . xxviii, 122, 590
  - as component of Unicode Standard .....59
  - changes .....56
  - properties in .....34
- Unicode character encoding model .....24, 31
  - see also* UTR #17, Unicode Character Encoding Model
- Unicode character literals
  - code point notation U+ .....583
- Unicode codespace
  - allocation numbers .....601
  - definition .....67
  - planes .....33
  - size .....1, 22
- Unicode Collation Algorithm (UCA) .....10
  - see also* UTS #10, Unicode Collation Algorithm
- Unicode Common Locale Data Repository (CLDR) . 590
- Unicode conferences .....589
- Unicode Consortium .....585
  - addresses .....591
  - Consortium membership in standards bodies 585
  - e-mail discussion list .....590
  - FTP site .....589
  - membership .....585
- Unicode data signature .....51, 561–562
- Unicode data types .....147–148
  - for C .....147–148
- Unicode encoding forms .....89–95
  - advantages of each .....28
  - conformance .....26, 62
  - definition .....90
  - fixed-width (UTF-32) .....26, 93
  - signatures .....562, 563
  - variable-width .....27, 93, 94
  - see also* encoding forms
- Unicode encoding schemes
  - conformance .....98–101
  - definition .....98
  - endian ordering .....30
  - see also* encoding schemes
- Unicode escape sequence notation \u1234 .....583
- Unicode Regular Expressions *see* UTS #18, Unicode Regular Expressions
- Unicode scalar values
  - definition .....89
- Unicode security mechanisms
  - see also* UTS #39, Unicode Security Mechanisms
  - Unicode security .....179
- Unicode Standard
  - allocation of encoded characters .....33–40
  - architecture .....7–10
  - areas .....34
  - benefits .....1
  - blocks .....34, 187
  - code charts .....571–580, 589
  - components .....59
  - conformance .....55–120
  - conformance of ISO/IEC 10646 implementations 600
    - corrections .....57
    - definitions for conformance .....65–69
    - design goals .....3
    - design principles .....10–19
    - errata .....57, 590
    - normative references to .....57, 63
    - number of characters .....2, 601
    - number of code points .....1, 22
    - script coverage .....2
    - security issues .....179
    - synchrony with ISO/IEC 10646 .....598
    - updates .....590
    - versions *see* versions of the Unicode Standard
    - see also* Version 6.2
- Unicode Standard Annexes (UAX) ..... xxvii, 586
  - as components of Unicode Standard .....59
  - conformance .....64
  - list of .....64
- Unicode string literals
  - code point notation \u1234 .....583
- Unicode strings .....32
  - definition .....91
- Unicode Technical Committee (UTC) .....585
- Unicode Technical Notes (UTN) .....589
- Unicode Technical Reports (UTR) .....586
  - abstracts .....587
- Unicode Technical Standards (UTS) ..... xxviii, 586
  - abstracts .....586
- policies .....590
- Web site .....589

- UnicodeData.txt ..... 115, 126
  - unification
    - as Unicode design principle ..... 17
    - see also* Han unification
  - Unified Repertoire and Ordering (URO) ... 415, 606
    - see also* Han unification
  - Unihan Database ..... 122, 418, 419, 578, 590, 607
  - Unihan.zip ..... 76, 122
  - unit separator (U+001F) ..... 545
  - Universal Character Set (UCS)
    - see* ISO/IEC 10646
  - universality
    - as Unicode design principle ..... 10
  - Unix
    - and UTFs ..... 29
    - newline function ..... 154
    - UTF-32 in ..... 27
    - UTF-8 in ..... 14
  - unsupported characters ..... 148–149
  - upadhmaniya ..... 295, 349
  - update version ..... 57
  - uppercase ..... 124, 172, 211
  - Uralic Phonetic Alphabet (UPA) ..... 203, 219
  - Urdu ..... 250
  - URO (Unified Repertoire and Ordering) ... 415, 606
    - see also* Han unification
  - UTF, Unicode Transformation Formats ..... 24, 90
    - advantages of each ..... 28
    - as encoding form or scheme ..... 100
    - binary comparison and sort order differences ..
      - 168, ..... 170
    - in APIs ..... 148
  - UTF-16 ..... 27, 93, 598
    - binary comparison and sort order caution ... 27
    - bit distribution (table) ..... 94
    - BOM in ..... 99, 561
    - encoding form (definition) ..... 93
    - encoding scheme (definition) ..... 99
    - encoding schemes ..... 30
    - in ISO/IEC 10646 ..... 598
    - in UTF-8 order ..... 171
    - surrogates and string handling ..... 32, 149
  - UTF-16BE (Big-endian) ..... 562
    - encoding scheme ..... 30
    - encoding scheme (definition) ..... 98
  - UTF-16LE (Little-endian) ..... 562
    - encoding scheme ..... 30
    - encoding scheme (definition) ..... 98
  - UTF-32 ..... 26, 93
    - BOM in ..... 100
    - encoding form (definition) ..... 93
    - encoding scheme (definition) ..... 100
    - encoding schemes ..... 30
    - in Unix ..... 27
  - UTF-32BE (Big-endian)
    - encoding scheme ..... 30
    - encoding scheme (definition) ..... 99
  - UTF-32LE (Little-endian)
    - encoding scheme ..... 30
    - encoding scheme (definition) ..... 99
  - UTF-8 ..... 27, 94, 598
    - ASCII transparency ..... 27
    - binary comparison and sort order ..... 29
    - bit distribution (table) ..... 94
    - BOM in ..... 98, 101, 561
    - byte ranges ..... 94
    - compared to multibyte encodings ..... 28
    - encoding form (definition) ..... 94
    - encoding scheme ..... 30
    - encoding scheme (definition) ..... 98
    - in Unix ..... 14
    - in UTF-16 order ..... 170
    - non-shortest form is invalid ..... 94, 179
    - preferred encoding for Internet protocols ..... 28
    - security and ..... 179
    - signature ..... 98, 101, 561
  - UTF-EBCDIC
    - see* UTR #16, UTF-EBCDIC
  - UTN (Unicode Technical Note) ..... 589
  - UTR (Unicode Technical Report) ..... 586
    - abstracts ..... 587
  - UTS (Unicode Technical Standard) ..... xxviii, 586
    - abstracts ..... 586
- ## V
- Vai ..... 454–455
    - reference materials ..... 637
  - valid (synonym for well-formed) ..... 92
  - variable-width Unicode encoding form ... 27, 93, 94
  - variants
    - compatibility ..... 20
    - fullwidth and halfwidth ..... 209
    - mathematical symbols ..... 516
    - small form ..... 209
    - standardized ..... 556
  - variation selectors ..... 142, 556
    - ideographic variation mark (U+303E) ..... 425
    - Mongolian free variation selectors ..... 444
  - variation sequences ..... 556
    - for Phags-pa ..... 340–341
  - Version 6.2 ..... 59
    - number of characters ..... 2, 601
  - versions of the Unicode Standard xxviii, 55, 590, 601–602
    - backward compatibility ..... 55
    - compared to ISO/IEC 10646 editions ..... 601
    - content ..... 56
    - interaction in implementations ..... 149
    - numbering ..... 56
    - property changes ..... 56
    - stability ..... 56
    - updates ..... 590
  - vertical tab (U+000B) ..... 154, 545
  - vertical text ..... 41, 192, 209
    - East Asian scripts ..... 406
    - Mongolian ..... 442
  - Vietnamese ..... 214, 220
    - ideographs ..... 405
  - virama ..... 189, 277
    - definition ..... 280
    - Kharoshthi ..... 358
    - Khmer ..... 376
    - Myanmar ..... 369
    - Philippine scripts ..... 392
    - virama-like characters ..... 142
  - visual order used for Thai and Lao ..... 16
  - vowel harmony
    - Mongolian ..... 445
  - vowel marks, Middle Eastern scripts ..... 245



vowel separator  
 Mongolian .....446

vowel signs  
 Indic .....43, 280  
 Khmer .....378  
 Philippine scripts .....392

## W

wchar\_t  
 and Unicode encoding forms .....28  
 in C language .....148

weak directional characters .....130

weather symbols .....524

Web site, Unicode Consortium .....589

Weierstrass elliptic function symbol .....497

well-formed  
 definition .....91

Welsh .....216

Where Is My Character? .....590

wide characters  
 data type in C .....148

wiggly fence (U+29DB) .....515

Windows newline function .....154

word breaks .....160, 546–548  
 in South Asian scripts .....366, 371, 383

word joiner (U+2060) .....546

writing direction *see* directionality

writing systems .....188–191

Wu (Shanghainese) .....413

## X

Xibe .....442

Xishuang Banna Dai .....384

XML  
*see* UTR #20, Unicode in XML and Other Markup Languages

## Y

yen currency sign .....495

Yi .....433–435  
 reference materials .....638

Yiddish .....246

Yijing Hexagram Symbols .....529

ypogegrammeni .....223

yuan currency sign .....495

## Z

Zapf Dingbats .....526

zero extension relation among encodings .....597

zero width joiner (U+200D) .....252, 549

zero width no-break space (U+FEFF) .....50, 63, 546  
 initial .....101, 562

zero width non-joiner (U+200C) .....252, 549

zero width space (U+200B) .....546  
 for word breaks in South Asian scripts . 366, 371, 383

zero-width space characters .....547

ZWJ *see* zero width joiner (U+200D)

ZWNBS *see* zero width no-break space (U+FEFF)

ZWNJ *see* zero width non-joiner (U+200C)

ZWSP *see* zero width space (U+200B)